图书情报工作
第62卷 第21期 2018年11月

词表语义组织研究的演进(1998－2018)*

陶俊
西北大学公共管理学院 西安 710127

摘要:[**目的/意义**]词表语义组织是馆藏语义化研究的重要组成部分,梳理该领域的历史演进有利于明晰重点并推动其更好发展。[**方法/过程**]在辨析词表语义组织领域核心术语的基础上,提出“标准规范——语义组织方法——支撑技术——词表应用”研究框架,基于该框架梳理中文叙词表语义研究代表文献。[**结果/结论**]提出词表语义组织的定义及其主体框架,揭示了词表、本体、关联数据等的核心概念及其有机联系;以叙词表为例梳理我国词表语义组织研究近10年来的代表性研究工作;比较词表传统研究与语义组织研究的内在异同,并对我国词表语义组织研究进行述评和展望。

关键词: 词表 关联数据 语义网 资源描述框架
分类号: G254.2
DOI: 10.13266/j.issn.0252－3116.2018.21.017

1 引言

关联数据云(LOD)显示,包含文献、生物、地理等多领域的RDF数据集正在激增,万维网已阔步向包含大量概念实体和实体间丰富语义关系的数据万维网演进。谷歌知识图谱(Google Knowledge Graph)、百度“知心”等语义搜索项目的推进使得语义网研究和实践正由传统学术领域的独奏曲向学术界与产业界共振的协作曲方向发展[1－2]。伴随着语义网的发展,知识组织正在历经传统工具辅助定位向数据智能研究的演变。尤其是关联数据运动开展以来,图书馆各类资源的语义化研究得到业内广泛关注,词表作为馆藏资源标引和辅助检索的基本工具,其语义组织研究是馆藏语义化研究的重要组成部分。

词表语义组织尚无统一的定义,本文将其界定为运用语义网相关标准与Web工程技术,推进词表在网络环境下的描述、关联和应用。作为一个交叉领域,前期相关研究有一些综述,例如:宋文等梳理了词表映射、互操作以及转换为本体的有关研究[3－5],薛春香等围绕词表互操作中的术语映射总结了基于词形、结构和语料的术语映射方法[6]。此外,还有一些学者从遗留资源关联数据发布[7]、术语服务和不同数据集的RDF关联等角度总结了代表性的项目、工具和实践[8－10]。与本主题密切相关的最新成果是M. L. Zeng等的*Knowledge organization system in the semantic web: a multi－dimensional review*一文,该文从关联开放数据集生产者、词表生产者和词表使用者等多维度探讨了关联开放数据集中词表的作用[11]。以上研究要么是围绕一个较小的领域开展综述,对词表语义组织的主题缺乏紧密关联;要么并不是围绕词表语义组织演进的逻辑框架来展开的;而且,由于词表语义组织横跨语义网和词表两个主题,以传统单一主题来梳理容易割裂交叉主题间的内在联系,导致相关工作在整体发展格局中的地位和作用无法知悉。为此,本文试图从更开放的层面辨析关键概念的演进及其内在异同,同时依托词表语义组织框架着重选取我国在词表语义组织领域的代表性研究工作进行综述。

本文的贡献主要表现在:①提出了词表语义组织的定义及其主体框架,揭示了词表、本体、关联数据等的核心概念及其有机联系;②以叙词表为例梳理了我国词表语义组织研究近10年来的代表性研究工作;③比较了词表传统研究与语义组织研究的内在异同,并对我国词表语义组织研究进行述评和展望。

2 概念辨析

词表语义组织研究近20年来在快速发展中,这不

* 本文系国家社会科学基金项目“多叙词表的开放关联与映射方法研究”(项目编号:14CTQ002)研究成果之一。
作者简介:陶俊(ORCID:0000－0002－5341－3509),讲师,博士,硕士生导师;E-mail:taoj@nwu.edu.cn。
收稿日期:2017－11－08 修回日期:2018－03－01 本文起止页码:140－148 本文责任编辑:易飞

仅体现在词表相关概念和范畴体系的拓展,同时,支持词表语义化的技术标准和方法如语义网、Web 工程和人工智能技术等也处于迭代更新之中,概念拓展和支撑技术的演进不仅增强了该领域的专业性,同时也容易造成术语或概念关系的混淆,进而使得领域研究趋缓。因此,厘清相关概念及其内在逻辑对于推动领域发展十分重要。

2.1 词表

词表有狭义与广义之分。狭义的词表指受控词表,又称叙词表,在我国通常指主题词表[12]。随着网络环境的发展,词表概念的内涵和外延在不断扩大。广义的词表包含了规范档、分类法、叙词表、语义网络和本体等类型(见图 1)。美国数字图书馆专家 G. Hodge 于 2000 年将其称为知识组织系统[13](knowledge organization system,简称 KOS)。知识组织系统概念的提出标志着传统分散的文献组织工具在网络环境下进入集约化发展的轨道。

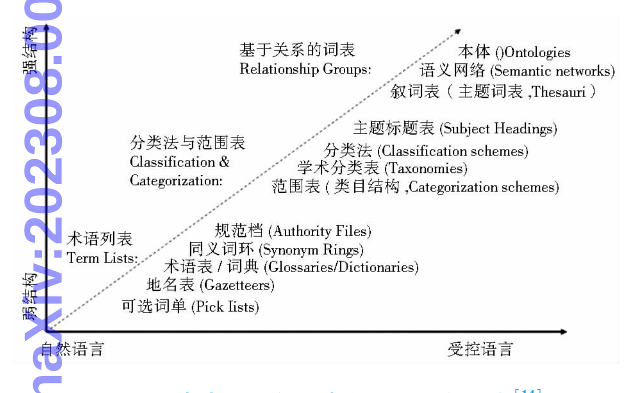


图 1 词表内涵由叙词表向两端不断扩大[14]

如前所述,词表的概念是在受控词表术语基础上发展起来的。受控词表包含了一系列术语并展示了不同的关系类型,其本质特征在于术语/概念及其词间关系的表达。2005 年,美国信息标准委员会发布的 ANSI/NISOZ39.19－2005 标准对受控词表的范围进行了拓展定义,根据受控的级别由弱到强将受控词表分为可选词单、同义词环、学术分类表和叙词表等类型。

2011 年,W3C 图书馆关联数据孵化小组将图书馆领域的关联数据集分为 RDF 元素集和值词表两种类型[15],它们也称为结构化词汇表和概念词表[16],前者通常有 MARC 元数据、DC、RDA 和 Bibframe,后者即是广义的词表,其核心是包含上位类关系、下位类关系和相关关系的叙词表。词表的外延相较过去得到了明显扩大,由原来的规范术语及术语关系拓展到表达某一领域的概念及其概念关系。主要词表及其语义表示标准规范如表 1 所示:

表 1 词表与语义表示标准规范

名称	标准规范中文名称	发布年度	标准组织
Z39.19－2005	单语种受控词表编制、格式与管理规则	2010	ANSI/NISO
BS8723	用于信息检索的结构化词表(Ⅰ－5)	2005－2008	BSI
ISO25964－1	用于信息检索的叙词表:单语种和多语种	2011	ISO
ISO25964－2	用于信息检索的叙词表:与其他词表的互操作	2013	ISO
XML	可扩展标记语言	1998	W3C
RDF	资源描述框架	1999	W3C
OWL1,2	Web 本体语言	2004 2008	W3C
SKOS	简单知识组织系统	2009	W3C

自计算机被应用于文献情报工作以来,词表描述格式由电子化逐步过渡到语义化。电子化的表示格式有数据库环境下的 MARC 和 Web 网页下的 HTML 格式,例如,AGROVOC 叙词表 2000 年开始由印本转为电子版存储于关系数据库中,我国于 2005 年实现了《中国分类主题词表》的 MARC 表示。事实上,HTML 格式只是将传统的文本电子化,不能适应计算机的语义表示和处理;而 MARC 格式虽能够通过元数据揭示语义信息,但其标准无法适应网络环境下数据开放共享和 Web 处理的需要。1998 年,XML 成为 Web 环境下数据表示的标准格式,它实现了将数据的语义信息(元数据)与数据内容相分离,同时能够满足网络环境交互和共享的需要,为此,美国国会图书馆等机构推动 MARC 格式向 MARCXML 和 MARC21 转变,各类元数据标准和词表也逐步采用 XML 语言作为数据表示的首选语言。为了适应网上资源的计算机智能处理,电子化的词表继续向实现语义表示的智能方面演化。

2.2 语义

词表语义化围绕语义网标准展开,本文将语义界定为包含概念关系逻辑和基于上述逻辑的形式化表示及其智能机制。语义网的核心是通过本体实现语义。本体包含两层含义:一是本体模型。即本体是将某一领域知识抽象后形成的概念及概念间关系的模型,通常描述了体系化的概念及其概念关系[17],本体模型概念的特性称为数据属性,概念间关系的特性称为对象属性。本体模型的原子概念陈述是以“主体－关系－客体”组成的 RDF 三元组;二是本体形式化。本体由本体模型走向计算机处理,需要一系列 Web 语言规范的支持。XML、RDF、OWL 是万维网联盟推出的 Web 数据的语义表示规范(见表 1)。基于上述语言可以对

本体模型的概念和属性构建多种不同的语义化描述。词表作为图书馆结构化资源中的一种类型,其语义表示方法与其他资源本质上相同,局部上略有差异。SKOS 是专门针对词表相对简化的结构提出来的表示规范,以区别于本体。我国于 2010 年实现了《中国分类主题词表》的 SKOS 表示[18]。

关联数据是由万维网之父蒂姆·伯纳斯李于 2006 年提出来的一项技术标准[19],其目标可概括为实现各类概念/实体及其关系的计算机建模、表示和关联发现。关联数据需要遵循两个基础标准:一是关联数据中的各类概念需要用 HTTP URI 表示,其目的是让每一个概念能够被 HTTP 协议访问,实现数据在 Web 环境下的开放共享;二是尽可能提供丰富的 URI 以发现或关联更多概念/实体。伴随越来越多单个关联数据集的开放发布,关联数据更重要的意义在于构建不同关联数据集相同或相关概念/实体的关联,owl:sameAs,rdfs:seeAlso 等关系词汇支撑概念实体的关联实现。

2.3 RDF、本体和关联数据的关系

RDF 是语义网数据表示的建模标准,本体和关联数据记录均需遵循 RDF 模型实现结构化。将某一数据资源生成关联数据需针对该资源体系建立本体模型并对本体模型形式化,关联数据本质是本体模型的实例。本体和关联数据的不同主要表现在语义建设机制的转变,具体可分为两个维度:①语义发展理念不同。语义网发展之初重在自主构建本体并通过更强大的推理逻辑来实现更丰富的语义;但关联数据提出后则侧重于重用包括词表在内的结构化资源来构建轻量级本体[20],同时依托不同资源的实体关联来实现语义的丰富化,淡化了原始本体构建及其复杂推理逻辑;②本体模型形式化理念不同。与传统本体模型重在自主定义单一的本体词汇集实现形式化不同,关联数据强调优先利用多个成熟的本体词汇集实现本体模型的形式化,通过最大化重用成熟本体词汇集有助于为后续与其他 RDF 数据集形成关联关系奠定基础。由于本体词汇集选取的差异,围绕某一数据资源建立的本体模型可能有多种不同的本体形式化方案。

3 研究内容

结合词表语义组织定义,本文将词表语义组织框架从标准规范到词表应用划分为 4 个层次(见图 2)。其基本逻辑是:词表语义组织是框架的核心,实现上述过程需以词表和 W3C 提供的相关标准作为基础,推进词表语义组织与应用的技术实现则需要 Web 工程技术的支持。图 2 中,词表语义描述、词表转化为本体、词表关联数据发布存在一定的包含关系,反映了词表语义组织在近 20 年演进中方法理念随着标准的动态变化而不断深化。

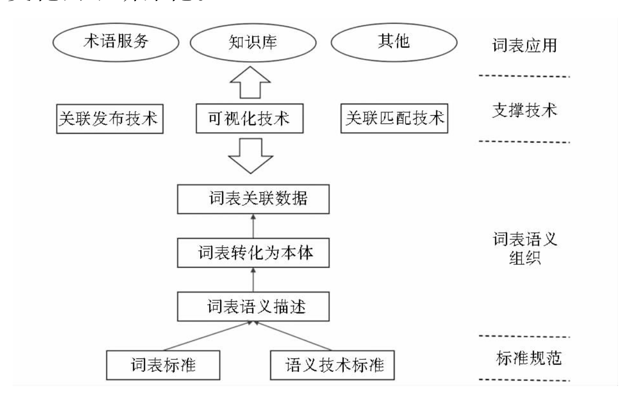


图 2　词表语义组织框架

图 2 中的支撑技术包括但不限于以上模块,其原因在于:①技术始终是为内容服务的,词表语义组织的内容伴随着技术的发展在不断丰富;②由于不同学者技术划分的差异及其内容侧重点不尽相同,支撑技术的表现形式可能多样。下文以图 2 框架对中文叙词表语义组织工作展开综述,重点在于反映主干工作的同时,揭示不同领域的有机联系。

3.1 词表语义组织

词表语义组织是在词表数据库化和 Web 网页化的基础上,面向语义网相关标准逐步发展起来的[21]。结合图 2,笔者将词表语义组织研究按照时代演进划分为"语义描述——本体转化——发布为关联数据"3 个层次(见表 2)。分类法、规范档和叙词表等的语义组织均包含以上过程,比较而言,叙词表在整个词表结构中处于关系层,相较其他更具代表性。下面重点以叙词表为例综述词表语义组织的代表工作,有关我国分类法和规范档的研究工作见文献[22-23]。

笔者首先揭示词表语义组织的演进逻辑,然后梳理我国词表语义组织的相关工作。

首先,从初期发展来看,由于语义相关标准尚不够成熟,图书馆学情报学界(简称"图情界")重在结合叙词表理解语义相关标准和概念[25]。作为在印本时代形成,以术语为中心,包含"用、代、属、分、参"等少数粗粒度词间关系的词表,与在网络环境下形成的以概念为中心、强调细粒度语义关系的本体有一定的近似关系。正因此,利用词表构建领域本体或者使词表实

陶俊. 词表语义组织研究的演进(1998 - 2018)[J]. 图书情报工作,2018,62(21):140 - 148.

表 2 词表语义组织演进

序号	组织层次	内容特征	备注
1	语义描述（2000 年至今）	结合 XML、RDF、OWL、SKOS 等表示规范探索描述方法	局部探索
2	本体转化（2003 年至今）	初期主要是利用叙词表构建本体,后来逐步专门针对叙词表实现本体形式化。一种思路是不改变词表结构实现形式化;一种思路是将粗粒度的词表层次结构改变为细化的网络结构,以概念为中心定义概念和属性并细化概念关系,在此基础上实现形式化描述,生成新的本体[24]	局部探索
3	发布为关联数据（2008 年至今）	以词表发布为中心,尽可能保留词表结构,基于关联数据标准制定 HTTP URI、定义本体模型和形式化方法作为描述词表关系的基础,构建批量转换程序,将新生成的资源运用关联数据工具发布	整体探索

现语义描述就成为图情界研究的中心议题。

第二,从研究对象来看,词表语义描述阶段因相关语义标准和技术不成熟,其研究以局部术语单元的理论探索为主;而在关联数据阶段,由于 SKOS 标准和关联数据发布技术相对成熟,其研究以词表整体探索为主,需以数据库版或 Web 版为基础。《农业科学叙词表》《中国分类主题词表》因在电子化建设方面相较其他词表先行一步,为词表的关联数据实践奠定了基础。第三,从语义组织层次来看,三者存在递进关系。语义描述重在以词表的语义化表示为中心,不涉及词表结构的调整;而将词表转换为本体则不仅涉及语义化表示,同时也会结合不同情境需求重新定义概念模型;词表发布为关联数据包含了以前语义描述、定义本体模型和本体形式化等过程,其不同在于融入了关联数据标准,同时侧重于将词表与其他书目资源一样作为具体的实例资源,这体现了词表作用在语义技术的应用下正在逐步跳出既有的工具辅助功能的定位,更多呈现出资源属性下数据智能的演化。

综合以上分析,词表语义组织是一个不断动态发展的过程。以此为基础,下文将综述中文叙词表语义转换的重点研究。曾新红、刘丽斌、段荣婷、鲜国建、刘华梅、欧石燕等均先后对中文叙词表的语义转换开展了方法探索,在业内具有代表性,如表 3 所示:

表 3 中文叙词表语义转换比较一览

主要研究者	标准类型	试验词表	转换内容	实现层次	转换语言	关联数据	年份
曾新红[26]	OWL	中分表	所有内容	方案描述	无	否	2005
刘丽斌[27]	SKOS	中分表	核心关系	方案描述 批量转换	Java	否	2009
段荣婷[28 - 29]	SKOS	中档表	所有内容	方案描述	无	否	2010
鲜国建[30 - 31]	SKOS	农表	核心关系	方案描述 批量转换	java	是	2013
刘华梅[32]	SKOS	中分表	所有内容	方案描述 批量转换	VB	否	2014
欧石燕[33]	SKOS	中分表	所有内容	方案描述 批量转换	Java	是	2015

注:《中分表》《中档表》《农表》依次指《中国分类主题词表》《中国档案主题词表》和《农业科学叙词表》

首先,从词表的语义表示语言和历史发展阶段来看,在未出现专门针对词表转换的标准规范 SKOS 以前,曾新红基于 OWL 语言对我国大型通用主题词表进行了语义描述探索[26]。词表类型集中在《中分表》《中档表》和《农表》等大型主题词表,《中分表》和《农表》实现了电子化,为语义化探索提供了坚实的基础。2009 年,SKOS 成为词表建设推荐标准后,后续研究主要以 SKOS 进行。

从转换内容来看,曾新红、刘丽斌和段荣婷等在词表转换研究上具有开拓性,曾新红和段荣婷主要结合 OWL 语言和 SKOS 语言对《中分表》和《中档表》提供了总体转换方案,包括对主表、附表、索引各部分实现 SKOS 描述[26,29]。刘丽斌等重在以词表中的“用、代、属、分、参”等核心关系为例进行了自动转换探索[27]。鲜国建等结合前述方法重点围绕《农表》进行了研究,同时构建了关联数据发布平台[31]。

从语义关系和技术实现视角来看,刘丽斌、鲜国建和欧石燕等运用高级编程语言 Java 实现了中文叙词表的 SKOS 自动转换。刘丽斌等最早开展中文叙词表语义转换并实现了《中分表》用、代、属、分、参、族等核心关系的自动转换[27]。鲜国建等采用 SKOS 和 SKOS - XL 实现了对《农业科学叙词表》的语义化表示并基于 Virtuoso 实现了关联数据发布[30 - 31]。欧石燕则在“用、代、属、分、参、族”等传统词间关系 SKOS 转换基础上,以 SKOS - EX 实现了组配、组面、族项等复杂概念的语义化表示,并通过 Java 语言实现了《中国分类主题词表》词表部分全描述和批量转换,并基于 Pubby 平台实现了关联数据发布[33]。刘华梅提出基于中分表 MARC 数据转换为 SKOS 的映射方案,同时采用 VB 语言实现了主题概念的批量转换[32]。

3.2 支撑技术研究

将词表发布为关联数据主要依托于 Web 工程技术,重点包括关联数据发布技术、可视化技术和关联匹配技术等。

3.2.1 关联发布 作为 RDF 数据集的一种类型,各类 RDF 数据集的关联发布方法均适用于词表[34-36]。词表 RDF 数据集的生成需要考虑词表建设基础和词表结构的差异,其步骤可简要概括为:HTTP URI 的确定——基于词表结构的本体建模——实体 RDF 化——实体关联化——RDF 文件——关联数据发布——开放 SPARQL 查询等阶段。从关系数据库文件输入输出角度来看,其输出包括文本文件——SQL 文件——RDF 文件(包括 rdf/xml,owl,skos 文件等)。如图 3 所示:

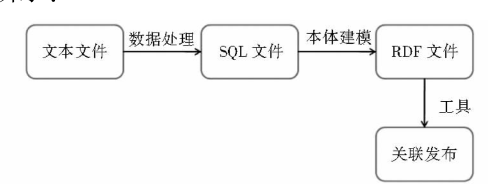


图 3 词表发布为关联数据的主要步骤

以中文叙词表的转换工作为例,鲜国建等[31]以《农表》的关系数据库为基础,结合《农表》的结构形成有关方案,包括设置 HTTP URI,“用、代、属、分、参”等主要关系与 SKOS 标签的对应等。刘丽斌与欧石燕等[27, 33]则以《中国分类主题词表》Web 版为基础,通过爬取其 HTML 格式得到大量词汇集合,经过预处理获得文本文件,然后将待转换的叙词表文本文件写入到 SQL 数据库,在此基础上确定本体建模方案,在 Jena API 等工具包的支持下编写 Java 转换程序实现批量转换得到 RDF 数据集。

3.2.2 可视化 RDF 数据集在反映细粒度概念以及多元语义关系方面具有优势,但其序列化格式因着眼于机器处理而难以供人们有效识别潜在关系,故可视化十分必要。本体开源可视化工具 WebVOWL、Protege、Welkin 等有利于揭示词表概念之间的语义关系。例如,范炜等在构建《中分表》主题词数据的术语服务原型系统中采用 Graphviz 和 Protovis 类库实现了关联数据的可视化[37]。洪娜等从开发平台、应用类型、开源与否、输入输出格式、是否支持三元组仓储、交互能力等方面比较了 RelFinder, Graphviz, RDF Gravity, RDFViz + +, Gruff 等 5 种可视化工具的差异[38]。一些较低门槛的工具更受大众欢迎,洪娜、任瑞娟、石泽顺等[39-41]以生物医学、中国知网和 LISTA 数据库的有关数据集为基础,利用在线工具构建 RelFinder 构建关联数据原型系统,通过可视化界面发现潜在关联关系。赵龙文和陈涛等则分别采用 Gruff 研究了政府领域关联数据的可视化和家谱关联数据实例中 RDF 数据的可视化[42-43]。

3.2.3 关联匹配 不同 RDF 数据集之间构建关联是关联数据五星标准的要求。截至 2017 年 2 月的 LOD 云图,仅有 10% 的数据成为五星关联数据[11]。继 2006 年开展中国《农业科学叙词表》和 AGROVOC 之间的映射[44],联合国粮农组织于 2011 年探索将 AGROVOC 与 EUROVOC、NALT、GEMET、STW、LCSH 和 RAMEAU 的关联匹配并将其发布为关联数据[45]。相比海外有关实践,我国词表研究在此领域的探索更少,大型词表间的关联研究几乎空白。陶俊[9]与朱雯晶等[9, 46]先后介绍了不同 RDF 数据集关联发现的自动化工具,包括 SILK 等。鲜国建等在探索《农表》关联数据发布中扼要介绍《农表》与 AGROVOC、NALT、LCSH 和 EUROVOC 的精确匹配关联结果,但并未结合实例阐述运用相关映射工具来实现关联匹配的实验过程[31]。相比词表关联,更多研究集中在书目数据集或文献资源数据集的关联上[47]。例如,虞为利等用海外书目数据片段同时融入 SILK 工具来查找等同关系探讨了书目数据集和 DBpedia 之间的关联[48],钟远薪和刘炜等结合上海图书馆书目数据和 DBpedia 数据探讨了不同数据集作者有关信息的关联[49]。

3.3 词表语义应用研究

我国词表语义应用方面的研究型探索主要体现在术语服务和语义知识库两方面。一方面,从图书情报角度来看,词表建设的目标是支撑网络环境下的术语服务,主要包含提供适用于人访问的 Web 界面术语查询以及支持计算机处理的应用编程接口; 另一方面,从语义网建设角度来看,词表发布为关联数据,本质上是一种精炼化的语义知识库,其概念关系可支持知识发现。从应用领域角度来看,术语服务和知识库可应用于医学、生物、法律等多领域场景。

3.3.1 术语服务 多位学者探讨了基于 REST 架构的术语服务技术实现。欧石燕等以《汉表》为例实现了 REST 架构的术语服务原型系统,同时从编目和元数据创建、信息检索和资源导航等情境阐释了术语服务应用形态[50-51]。曾新红等基于分类法系统 CLSS 提供了基于 Web 服务 API 和 Web 页面检索的术语服务[52-53]。此外,一些学者以相关 SKOS 文档或 OWL 文档为基础构建术语服务原型系统。徐雷和董慧以美国

陶俊. 词表语义组织研究的演进(1998－2018)[J]. 图书情报工作,2018,62(21):140－148.

国家癌症研究所的 NCI 癌症叙词表 OWL 文档作为数据源,依托图形数据库 Neo4j 作为存储平台构建了一个 REST 架构的术语服务[54]。范炜等基于 Cherrypy + TDB + Joseki 作为关联数据发布框架,在此基础上构建了术语服务原型[37]。

3.3.2 知识库 部分学者围绕语义知识库构建探索了词表在支持概念检索方面的作用。北京大学王军以《中分表》的类目和主题词以及元数据为基础,构建书目本体模型 KVision 并形式化,同时以北京大学图书馆计算机领域的书目数据作为本体实例构建语义知识库进而实现概念[55]。欧石燕等以多个文献数据源的关联数据转换为中心,实现上述不同数据源中受控词汇、人名、地名等的 RDF 关联[47];在此基础上,进一步运用自然语言处理技术探索将自然语言转化为结构化的 SPARQL 查询技术,实现对多个 RDF 数据集成搜索和自动问答[56]。

4 讨论

以上从 4 个方面梳理了我国词表语义组织领域的代表性研究,同时阐释了各模块的作用以及模块间的关联递进关系,下面从科学问题和研究特点两方面进一步讨论。

4.1 词表语义组织问题的内在联系

传统图书情报领域,词表研究的三大典型问题是词表标准规范、词表构建和词表的术语映射(见表 4)。语义网标准下则更侧重于探究“词表语义描述——词表转化为本体/关联数据发布——不同数据集间开放关联”等问题。这些新问题均是建立在前述经典问题基础之上。首先,从术语维度看,概念和概念关系的语义表示尽管更多依赖于各类语义规范的应用,但同时离不开对词表标准的深入理解;其次,从输出维度看,词表转化为本体以及关联数据发布本质上是促进计算机处理,它需要以丰富的词汇和词间关系作为基础,词表构建与更新的目标正是实现上述目标[57];再次,从关系维度看,不同数据集间开放关联的重点是探讨与其他 RDF 数据集间的概念等同关系,这与词表映射探讨术语或概念的映射具有一致性。尽管在技术实现上可能有 Silk 等针对关联开放数据集的关联发现方法,但从底层的映射或对齐方法来讲,大多仍以各类字符串相似度算法及其 API 来实现[9],这与本体匹配的相关方法具有一致性。总之,词表语义组织和词表本身建设是互为联系的整体。

表 4 词表语义组织研究相关问题比较

比较维度	词表本身	语义网标准
基础维	标准规范研究	词表概念和属性的语义表示
目标维	词表构建与更新	词表转化为本体/关联数据发布
关系维	词表映射/互操作	不同数据集间关联匹配

4.2 研究实践性强,专业化研究队伍较少

总体来看,我国词表研究呈现出两大特点:

(1)研究实践性强。词表研究以国家科技文献中心支撑机构为主,侧重于工程实践,比如,中国科学院文献情报中心围绕面向外文知识组织平台与集成系统建设开展了大量工作[58-59];中国科学技术信息研究所以《汉语主题词表》的更新为基础,深入推进国家叙词库建设,形成了较强的实践特色[60];中国医学科学院长期围绕医学领域开展词表的语义应用研究跟踪和实践探索;中国农业科学院在农业科学词表关联发布与平台方面形成了一定特色[31];在业内形成了一定影响;此外,国家图书馆与全国图书情报领域多家单位围绕《中国分类主题词表》的开放性研究形成了影响。相比上述方面,更多机构主要以自由探索为主,缺乏集聚性。总体来讲,综合型的实践跟踪和概要研究居多,围绕具体科学问题的工程研究和实践创新相对缺乏,体现出普及性和碎片化的特点。

(2)专业化研究队伍较少。相比其他研究,图书馆实践部门和高校图情学者对词表的连续研究规模相对不足。笔者以为,这一现象具有内外两方面的因素。内因方面,词表语义组织研究的专业门槛在逐步提升。伴随词表标准适应网络环境,词表研究在变革中形成大量新的概念术语,与此同时,语义网、Web 工程和自然语言处理等新技术正日益主导词表建设的发展,学科交融使得词表研究不再属于传统图书情报领域知识范畴,而更多依赖于计算机应用的支撑;外因方面,作为图书情报领域的核心领域,词表的传统固有定位在网络环境下逐步边缘化。以上两方面使得持续开展词表探索的人员正在日益减少,词表领域正在随着人工智能技术的应用和语义搜索服务的形成发生潜在变化,这给词表语义组织研究带来了挑战的同时也意味着机遇。

5 研究展望

词表语义组织是适应网络时代需要逐步发展起来的。伴随词表形成 RDF 数据集并以关联数据发布,其

应用不再拘泥于文献检索和辅助语义标引等术语服务,搜索引擎在向智能化检索潜在转型的今天仍将是网络化生存的基础工具,基于此,未来在优化术语服务实践的同时,围绕语义搜索探究词表应用是深化研究的重点。具体而言,可从三方面展望:

(1)加强大型词表语义关联匹配探索。伴随语义环境的形成,语义搜索的发展方向之一是向信息关联方向发展,对《汉语主题词表》和《中分表》等大型词表与国内外其他词表开展对照映射[18,61],推进《中分表》的分面化改造等,有利于支持基于信息发现的深度检索,进而提升大型词表在新时代的应用价值。在此基础上,进一步基于关联匹配,探索不同类型词表关联方法与实践,尤其是词表与各类 RDF 数据集(词表数据集、其他资源数据集)间的开放关联实验和技术创新研究是未来的重点[62]。

(2)拓宽词表语义组织的技术范畴[63]。当前图情界的词表语义组织技术多集中在 W3C 倡导的关联数据技术。从发展趋势看,关联数据相关标准和技术只是其中的一个分支,与此对应的是,人工神经网络模型、认知计算等得到快速发展并在多领域应用使得上述方法成为发展数据智能的利器。显然,利用本体、元数据等通过人为建立知识表达模型的语义网方法正受到来自以词、句向量为基础实现全程无人工干预的智能计算方法的挑战。正如艾思维尔首席架构师 B. P. Allen 所指出的,语义网是基于人而不是基于机器,它在帮助机器学习怎样阅读方面有不足,未来需要使用机器阅读来建立知识图谱[64]。因此,以相关平台系统或 API 应用为基础,吸收数据库、自然语言处理、机器学习(深度学习)等多学科领域探索知识图谱的有关方法实践是未来深化词表语义组织研究的重要手段[65-66]。

(3)深化词表语义组织的应用领域。首先,进一步推进各类资源的关联数据发布是深化语义搜索的基础。通过构建词表与领域数据集的 RDF 链接使词表概念成为各类领域数据集的聚合中介。换言之,领域数据集关联发布越多,未来词表数据集潜在聚合面越广。英国、芬兰等以数字人文运动为纽带,进一步推进历史、法律等资源的关联发布是上述工作的体现。其次,结合元数据词表探索融入商业、社交网络和智慧交通等更广泛的人物、地理和应用场景[67-68],借助词表语义标引研究大数据环境下的用户画像和个性化推荐与挖掘是拓展词表应用的重要方面。

6 结语

自 1998 年 NKOS(network knowledge organization system,简写 NKOS)小组成立以来,词表由网络化向语义化不断演进,词表词义组织研究在概念、内容、方法与技术等方面均有了长足发展。相比传统综述研究围绕单一的词表或语义网等具体领域进行梳理,本文选择将词表与语义网等标准相结合并从纵向勾勒词表语义组织的整体发展及其内在机制,以弥补历史同类研究的不足。当然,本文也存在不足之处。首先,本文的内容跨度较大,致使在突出重点的同时横向局部的分析上稍显不足,例如有关词表标准的研究和词表语义化的领域应用等介绍较少;其次,本文在词表语义组织研究工作方面主要围绕中文词表语义化工作进行,海外同类研究对于深化词表研究同样重要。以上不足之处将在后续研究中加以弥补。

参考文献:

[1] 王昊奋. 大规模知识图谱技术[J]. 中国计算机学会通讯, 2014, 10(3):64-68.
[2] 邹磊. 知识图谱的数据应用和研究动态[J]. 中国计算机学会通讯,2017, 13(8):49-54.
[3] 陈辰,宋文. 叙词表映射研究综述[J]. 图书情报工作,2012, 56(12):113-119.
[4] 段瑞龙,宋文. 国内外叙词表转换本体方法研究综述[J]. 情报杂志,2012, 31(7):66-71.
[5] 宋文. 知识组织体系语义互操作研究[J]. 图书馆论坛,2012(11):117-121.
[6] 薛春香,乔晓东,朱礼军. KOS 互操作中的术语映射研究综述[J]. 现代图书情报技术,2010, 26(2):31-37.
[7] MARJIT U,SHARMA K,SARKAR A, et al. Publishing legacy data as linked data: a state of the art survey[J]. Library Hi Tech, 2013, 31(3):520-535.
[8] 欧石燕. 国外术语注册与术语服务综述[J]. 中国图书馆学报, 2014,40 (5):110-126.
[9] 陶俊. 基于 Linked Data 的 RDF 关联框架综析[J]. 现代图书情报技术,2011(12):1-8.
[10] GOLUB K,TUDHOPE D,ZENG M L, et al. Terminology registries for knowledge organization systems: functionality, use, and attributes[J]. Journal of the Association for Information Science and Technology,2014, 65(9):1901-1916.
[11] ZENG M L,PHILIPP M. Knowledge organization system(KOS) in the semantic Web: a multi-dimensional review [EB/OL]. [2018-05-05]. https://arxiv.org/abs/1801.04479.
[12] 戴维民. 信息组织[M]. 2 版. 北京:高等教育出版社,2009:

133.
[13] HODGE G. Systems of knowledge organization for digital libraries: beyond traditional authority files[R/OL]. [2017 -12 -20]. https://www.clir.org/wp-content/uploads/sites/6/pub91.pdf.
[14] ZENG M L,FAN W. SKOS and its application in transferring traditional thesauri into networked knowledge organization systems[EB/OL]. [2017 -10 -18]. https://www.researchgate.net/publication/265817740_SKOS_and_Its_Application_in_Transferring_Traditional_Thesauri_into_Networked_Knowledge_Organization_Systems.
[15] W3C. Library data resources[EB/OL]. [2017 -10 -15]. https://www.w3.org/2001/sw/wiki/LLD/Library_Data_Resources,2011.
[16] NISO. Vocabulary management[EB/OL]. [2018 -02 -20]. http://www.niso.org/standards-committees/vocab-mgmt.
[17] BAWDEN D,ROBINSON L. Introduction to Information Science [M]. London:Facet Publishing, 2012: 114.
[18] 范炜. 走向开放化、语义化与关联化的《中国分类主题词表》[J]. 图书馆学与资讯科学(台湾),2017, 43(1):155 -170.
[19] BIZER C,HEATH T,BERNERS-LEE T. Linked Data-the story so far[J]. International journal semantic Web information system, 2009, 5(3):1 -22.
[20] 纪姗姗,刘峥,宋文. 叙词表向本体重构的关键技术研究[J]. 图书与情报,2013(1):8 -12.
[21] 王军,张丽. 网络知识组织系统的研究现状和发展趋势[J]. 中国图书馆学报,2008, 34(1):65 -69.
[22] 张士男,宋文.《科图法》SKOS 描述方案设计[J]. 现代图书情报技术,2010(6):7 -11.
[23] 赵捷,贾君枝. 数据网络中中文名称规范档的建设与发展[J]. 图书情报工作,2017, 61(22):134 -139.
[24] 贾君枝.《汉语主题词表》转换为本体的思考[J]. 中国图书馆学报,2007, 33(4):41 -44
[25] 毛军. 基于 RDF 的叙词表研究[J]. 情报学报,2003, 22(2):163 -168.
[26] 曾新红.《中国分类主题词表》的 OWL 表示及其语义深层揭示研究[J]. 情报学报,2005, 24(2):151 -160.
[27] 刘丽斌,张寿华,濮德敏,等.《中国分类主题词表》的 SKOS 描述自动转换研究[J]. 中国图书馆学报,2009, 35(6):56 -60.
[28] 段荣婷.《中国档案主题词表》语义网络化应用研究[J]. 档案学研究,2010(6):66 -70
[29] 段荣婷. 基于简约知识组织系统的主题词表语义网络化研究——以《中国档案主题词表》为例[J]. 中国图书馆学报, 2011, 37(3):54 -65.
[30] 鲜国建,赵瑞雪,朱亮,等. 农业科学叙词表的 SKOS 转化及其应用研究[J]. 现代图书情报技术,2012(10):16 -20.
[31] 鲜国建,赵瑞雪,寇远涛,等. 农业科学叙词表关联数据构建研究与实践[J]. 现代图书情报技术,2013 (11):8 -14.
[32] 刘华梅.《中国分类主题词表》主题词 SKOS 化描述及自动转换研究[J]. 图书馆建设,2014 (8):29 -32,36.
[33] 欧石燕. 中文叙词表的语义化转换[J]. 图书情报工作,2015, 59(16):110 -118.
[34] RADULOVIC F,POVEDA-VILLALON M,DANIELVILA-SUERO, et al. Guidelines for Linked Data generation and publication: an example in building energy consumption[J]. Automation in construction,2015, 57:178 -187.
[35] 夏翠娟,刘炜,赵亮,等. 关联数据发布技术及其实现——以 Drupal 为例[J]. 中国图书馆学报,2012,38 (1):49 -57.
[36] 沈志宏. 关联数据发布流程与关键问题研究——以科技文献、科学数据发布为例[J]. 中国图书馆学报,2013, 39(2):53 -62.
[37] 范炜,邹庆.《中国分类主题词表》的术语网络服务探索——以主题词规范数据为例[J]. 图书情报工作,2012, 56(14):40 -46.
[38] 洪娜,钱庆,范炜,等. 关联数据中关系发现的可视化实践[J]. 现代图书情报技术,2013 (2):11 -17.
[39] 洪娜,朱凯,王军辉, 等. 利用 RelFinder 实现生物医学语义关系发现 [J]. 情报杂志,2013 (4):142 -148.
[40] 任瑞娟,濮德敏,张媛. 基于五维学术关系发现的知识脉络可视化实践[J]. 大学图书馆学报,2016 (1):69 -75.
[41] 石泽顺,肖明. 基于 RelFinder 的图情学科关联数据语义关系发现实践[J]. 图书情报工作,2017, 61(17):139 -148.
[42] 赵龙文,罗力舒. 基于关联数据的政府数据开放:模式、方法与实现[J]. 图书情报工作,2017, 61(19):102 -112.
[43] 陈涛,夏翠娟,刘炜,等. 关联数据的可视化技术研究与实现[J]. 图书情报工作,2015, 59(17):113 -119.
[44] LIANG A C,SINI M. Mapping AGROVOC and the Chinese agricultural thesaurus: definitions, tools, procedures[J]. New review of hypermedia and multimedia,2006, 12(1):51 -62.
[45] MORSHED A,CARACCIOLO C,JOHANNSEN G, et al. Thesaurus alignment for linked data publishing [C]//Proceedings of the 2011 international conference on Dublin core and metadata applications. Hague: The National Library of the Netherlands, 2011.
[46] 朱雯晶,夏翠娟,刘炜. SILK 关联发现框架综析[J]. 现代图书情报技术,2013 (4):18 -24.
[47] 欧石燕,胡珊,张帅. 本体与关联数据驱动的图书馆信息资源语义整合方法及其测评[J]. 图书情报工作,2014, 58(2):5 -13.
[48] 虞为,陈俊鹏. 基于 MapReduce 的书目数据关联匹配研究[J]. 现代图书情报技术,2013 (9):15 -22.
[49] 钟远薪,李田章,刘炜. OPAC 混搭关联数据应用研究[J]. 现代图书情报技术,2013(4):25 -29.
[50] 欧石燕. 基于 SOA 架构的术语注册和服务系统设计与应用 [J]. 中国图书馆学报,2011,37 (5):13 -25.
[51] 欧石燕,唐振贵,苏翡斐. 面向信息检索的术语服务构建与应

用研究[J]. 中国图书馆学报,2016,42 (2):32 -51.
[52] 曾新红,黄华军,刘春燕, 等. ISO 5127 的 SKOS 语义描述方案及其共享服务系统研究[J]. 2017, 61(21):123 -129.
[53] 黄华军,曾新红,林伟明,等. 中文知识组织系统形式化语义描述标准体系研究(二)——分类法共享服务系统 CLSS 研究与实现[J]. 中国图书馆学报,2015,41 (2):17 -28.
[54] 徐雷,董慧. 基于 REST 架构的术语注册与服务研究实现[J]. 现代图书情报技术,2012 (7/8):59 -65.
[55] 王军. 基于传统知识组织资源的本体自动构建[J]. 情报学报,2009, 28(5):651 -657.
[56] 欧石燕,唐振贵. 面向图书馆关联数据的自动问答技术研究[J]. 中国图书馆学报,2015,41 (6):44 -60.
[57] 常春. 网络环境下叙词表的编制与发展[M]. 北京:科学技术文献出版社,2015.
[58] 孙坦,刘峥. 面向外文科技文献信息的知识组织体系建设思路[J]. 情报杂志,2013 (1):2 -7.
[59] 刘峥,纪姗姗. 叙词表标准的数据模型研究[J]. 图书情报工作,2013, 57(2):103 -108.
[60] 吴雯娜,鲍秀林. 国家叙词库的体系结构与数据模型[J]. 中国图书馆学报,2016,42 (2):81 -96.
[61] 鲍秀林,吴雯娜. 语义映射质量及影响因素分析[J]. 中国图书馆学报,2016, 42(5):57 -67.
[62] BINDING C,TUDHOPE D. Improving interoperability using vocabulary linked data [J]. International journal of digital libraries, 2016, 17(1):5 -21.
[63] ZENG M L. Smart data for digital humanities[J]. Journal of data and information science,2017, 2(1):1 -12.
[64] ALLEN B P. The role of metadata in the second machine age[EB/OL]. [2018 - 08 - 10]. http://dcevents. dublincore. org/IntConf/dc -2016/paper/view/464/534.
[65] 李芳,刘胜宇,刘峥. 生物医学语义关系抽取方法综述[J]. 图书馆论坛,2017 (6):61 -69.
[66] CHOUDHURY S. The role of metadata in an open knowledge age? [EB/OL]. [2018 -08 -10]. http://dcevents. dublincore. org/IntConf/dc -2017/paper/viewFile/526/648.
[67] 王军,张璐,张文军. 基于用户需求的电商导购机制设计[J]. 情报学报,2016, 35(7):730 -738.
[68] GRACY K F,ZENG M L,SKIRVIN L. Exploring methods to improve access to music resources by aligning library data with linked data: a report of methodologies and preliminary findings[J]. Journal of the American Society for Information Science and Technology,2013, 64(10):2078 -2099.

A Survey for Research on Semantic Organization of Vocabularies(1998 -2018)

Tao Jun
School of Public Management, Northwest University, Xi'an 710127

Abstract: [**Purpose/significance**] Semantic organization of vocabulary, an important part in collection semantic research, is the focus of knowledge organization study. A research review in this field is helpful to promote its development. [**Method/process**] Based on the analysis of the core terms in the field of semantic organization of vocabulary, this paper proposes the analytical framework of "standard specification-semantic organization method-supporting technology-vocabulary application". With above framework, the paper reviews literature about method, technology and application. [**Result/conclusion**] Firstly, the paper gives the definition and main frame of vocabulary semantic organization, discusses the core concepts and their relationship including vocabulary, ontology and linked data. Then taking the example of thesaurus, it summarizes the typical research of vocabulary semantic organization in China in recent ten years. And it compares the traditional vocabulary research and semantic research. On the basis of summarizing the above literature, the current situation and future development of semantic organization of Chinese vocabulary are discussed.

Keywords: vocabulary linked data semantic Web resource descriptive framework